Likelihood: Efficiency

Patrick Breheny

October 20

Introduction

- Today we will prove the information inequality, which establishes a lower bound on the variability of an estimator in terms of the Fisher information
- This leads us to the idea of an "efficient" estimator, as any estimator that achieves this lower bound is, in some sense, optimal

Information inequality: 1D

- First, let's take a look at the information inequality in the case of a scalar estimator
- **Theorem (Information inequality):** Let $\hat{\theta}$ be an estimate of θ, and let $g(\theta) = \mathbb{E}\hat{\theta}$ be finite. Suppose $X \sim p(\cdot|\theta^*)$ and $d/d\theta$ can be passed under the integral sign with respect to both $\int dP$ and $\int \hat{\theta} dP$. Finally, suppose $\mathscr{I}(\theta) > 0$ for all θ. Then

$$\mathbb{V}\hat{\theta} \geq \frac{\dot{g}(\theta^*)^2}{\mathscr{I}(\theta^*)}$$

Remarks

- The preceding theorem is somewhat vague about whether X is a single observation, a random sample, iid ...
- The reason is that it applies to all of these situations; just keep in mind that in the case of a random sample $X_1, X_2, \ldots, X_n$, the derivative must be able to be passed inside the joint distribution of all the X's
- Accordingly, please note that this is *not* an asymptotic theorem – it is an inequality that is true for all values of n

Attainment

- Is it possible for estimators to achieve this bound? (i.e., to have the minimum possible variance?)
- An interesting theorem due to Wijsman (1973) is that equality is only possible in the information inequality if $\hat{\theta}$ is linearly related to the score
- In other words, the only situation in which the lower bound is attainable (for all θ, for all n) is when $\hat{\theta}$ is the sufficient statistic of an exponential family

Cramér-Rao lower bound

- The information inequality is often restated in terms of the bias of an estimator
- Letting $b(\theta) = g(\theta) - \theta$ denote the bias of $\hat{\theta}$, and assuming we have an iid sample, then the information inequality becomes

$$\mathbb{V}\hat{\theta} \geq \frac{(1 + \dot{b}(\theta^*))^2}{n\mathscr{I}(\theta^*)}$$

  or, in the case of an unbiased estimator,

$$\mathbb{V}\hat{\theta} \geq \frac{1}{n\mathscr{I}(\theta^*)}$$

- In this form, the inequality is known as the *Cramér-Rao lower bound*

Remarks

- Recall that the mean squared error of an estimator is

$$\begin{aligned}\text{MSE} &= \mathbb{E}\{(\hat{\theta} - \theta^*)^2\} \\ &= \text{Bias}^2 + \text{Var}\end{aligned}$$

- Thus, among unbiased estimators, the CRLB represents the minimum possible MSE
- However, this requirement is rather artificial: it is often the case that biased estimators can be constructed with a lower MSE than the best unbiased estimator

Example #1

- The CRLB is not always attainable
- For example, if $X_i \overset{\text{iid}}{\sim} \mathrm{N}(\mu, \sigma^2)$, the CRLB for σ^2 is $2\sigma^4/n$
- It turns out that this bound is unobtainable if μ is unknown; all unbiased estimators have a higher variance than this
- For example, letting s^2 represent the usual unbiased estimator of the variance,

$$\mathbb{V}s^2 = \frac{2\sigma^4}{n-1} > \frac{2\sigma^4}{n}$$

Example #2

- Keep in mind also that the CRLB only applies when we can pass the derivative under the integral
- One common model for which this cannot be done is $X_i \overset{\text{iid}}{\sim} \text{Unif}(0, \theta)$
- In this case, one might think that the CRLB is θ^2/n
- However, $\hat{\theta} = (n+1)X_{(n)}/n$ is an unbiased estimate of θ with

$$\mathbb{V}\hat{\theta} = \frac{\theta^2}{n(n+2)} < \frac{\theta^2}{n}$$

- The "real" CRLB here is not well defined

Information inequality: Multiparameter

- Now, let's prove the information inequality for the case of a vector of parameters
- **Theorem (Information inequality):** Suppose $X \sim p(x|\boldsymbol{\theta}^*)$, with $\mathscr{I}(\boldsymbol{\theta}^*)$ positive definite. Let $\hat{\gamma}$ be an estimator with finite expected value $g(\boldsymbol{\theta})$. If $\nabla^2_{\boldsymbol{\theta}} f(x|\boldsymbol{\theta}^*)$ exists and can be passed under the integral sign with respect to $\int dP$ and $\int \hat{\gamma} dP$, then

$$\mathbb{V}\hat{\gamma} \succeq \nabla g(\boldsymbol{\theta}^*)^\top \mathscr{I}(\boldsymbol{\theta}^*)^{-1} \nabla g(\boldsymbol{\theta}^*)$$

- Recall that $\mathbf{A} \succeq \mathbf{B}$ means $\mathbf{A} - \mathbf{B}$ is positive semidefinite

Special case: $g(\boldsymbol{\theta}) = \boldsymbol{\theta}$

- In the special case where we have iid data and an unbiased estimator of $\boldsymbol{\theta}$, we have the simple result that:

$$\mathbb{V}\hat{\boldsymbol{\theta}} \succeq \frac{1}{n}\boldsymbol{\mathscr{I}}(\boldsymbol{\theta}^*)^{-1},$$

  the Cramér-Rao lower bound in d dimensions
- A related case: suppose we are estimating only a subset of $\boldsymbol{\theta}$, say, $\boldsymbol{\theta}_1$, with remaining parameters so-called "nuisance parameters"
- What is the impact on the CRLB?

Nuisance parameters

- A common notation convention when dealing with partitions of the information matrix is to let $\mathscr{I}_{11}$ denote the $(1,1)$ block of the information matrix, and $\mathscr{I}^{11}$ denote the $(1,1)$ block of $\mathscr{I}^{-1}$ (and so on for other partitions, and for the observed information)
- Using this notation, the CRLB for estimating $\boldsymbol{\theta}_1$ is $\mathscr{I}^{11}/n$, as opposed to the CRLB for estimating $\boldsymbol{\theta}_1$ in the case where $\boldsymbol{\theta}_2$ is known: $\mathscr{I}_{11}^{-1}/n$
- Personally, I don't like this notation and prefer $\mathscr{V}$ to denote $\mathscr{I}^{-1}$ and $\mathbfcal{V}$ to denote $\mathbfcal{I}^{-1}$, mainly because $\mathscr{I}^{11}$ tends to cause some confusion as looking like an information, when it is very much *not* an information of any kind

Information loss due to nuisance parameters

- Recall that the relationship between these two quantities is given by the Schur complement, which we restate here in terms of our new information matrix notation (for the sake of compactness, I'm suppressing the dependence on $\boldsymbol{\theta}$ here):

$$\mathscr{V}_{11}^{-1} = \mathscr{I}_{11} - \mathscr{I}_{12}\mathscr{I}_{22}^{-1}\mathscr{I}_{21},$$

or, if you prefer the superscript notation,

$$(\mathscr{I}^{11})^{-1} = \mathscr{I}_{11} - \mathscr{I}_{12}\mathscr{I}_{22}^{-1}\mathscr{I}_{21};$$

recall that $\mathscr{I}_{22}^{-1}$ is positive definite, so the term being subtracted cannot be negative ($\mathscr{I}_{11} \succeq \mathscr{V}_{11}^{-1}$)
- In other words, $\mathscr{I}_{12}\mathscr{I}_{22}^{-1}\mathscr{I}_{21}$ is the cost of not knowing $\boldsymbol{\theta}_2$ when estimating $\boldsymbol{\theta}_1$ (i.e., the information we've lost)

Orthogonality

- Only if $\mathscr{I}_{12} = \mathbf{0}$ do we suffer no information loss
- This can indeed happen; when it does, the parameters $\boldsymbol{\theta}_1$ and $\boldsymbol{\theta}_2$ are said to be *orthogonal*
- For example, consider the case where $X_i \overset{\text{iid}}{\sim} \mathrm{N}(\mu, \sigma^2)$
- Here, $\bar{x}$ is unbiased for μ and achieves the CRLB regardless of whether we know σ^2 or not
- Such situations, however, are more the exception than the rule

Efficiency

- The information inequality and CRLB are of somewhat limited use in finite samples, since they are only achieved in special cases
- Reaching the CRLB *asymptotically*, on the other hand, is a different matter, and a much more attainable goal for a hardworking little estimator
- **Definition:** Let $X_i \overset{\text{iid}}{\sim} p(x|\boldsymbol{\theta}^*)$. Suppose a sequence of estimates $\hat{\boldsymbol{\theta}}_n$ for $\boldsymbol{\theta}$ satisfies $\sqrt{n}(\hat{\boldsymbol{\theta}}_n - \boldsymbol{\theta}) \xrightarrow{\text{d}} \text{N}(\mathbf{0}, \boldsymbol{\Sigma}(\boldsymbol{\theta}))$. The sequence is said to be *asymptotically efficient* if $\boldsymbol{\Sigma}(\boldsymbol{\theta}) = \boldsymbol{\mathscr{I}}^{-1}(\boldsymbol{\theta})$ for all $\boldsymbol{\theta}$.
- While "asymptotically efficient" is a more accurate term, it is common to refer to such estimators as "efficient"

Efficiency and maximum likelihood

- As we have already shown, the MLE is asymptotically efficient (under certain regularity conditions)
- Thus, the MLE is in some sense optimal: at least asymptotically, no sequence of unbiased estimators can improve upon the MLE's accuracy
- For a long time in statistics, it was thought that no biased estimators could do better either; this belief, however, was upended by JL Hodges

Superefficiency

- Suppose $X_i \overset{\text{iid}}{\sim} \mathrm{N}(\theta, 1)$ so that $\sqrt{n}(\hat{\theta} - \theta) \sim \mathrm{N}(0, 1)$
- Consider the biased estimator

$$\tilde{\theta} = \begin{cases} 0 & \text{if } |\hat{\theta}| < n^{-1/4} \\ \hat{\theta} & \text{if } |\hat{\theta}| \geq n^{-1/4} \end{cases}$$

- Now, $\mathbb{P}\{|\hat{\theta}| < n^{-1/4}\} \to 1$ if $\theta = 0$ and $\to 0$ otherwise
- Thus, $\sqrt{n}(\tilde{\theta} - \theta) \xrightarrow{\mathrm{d}} \mathrm{N}(0, v)$, where $v = 1$ if $\theta \neq 0$ and $v = 0$ if $\theta = 0$

Superefficiency (cont'd)

- In other words, v improves upon the CRLB; a so-called "superefficient" estimator
- It's a pretty neat counterexample, although not necessarily a serious challenge to likelihood theory, as it can be shown (Le Cam, 1952) that the set of superefficient points always has Lebesgue measure zero
- This is sort of like saying that the MLE achieves the optimal variance almost everywhere, but this would only be a meaningful statement with a Bayesian prior, as otherwise there is no probability distribution associated with θ

Two Cauchy estimators

- To get a sense of why efficiency is a useful concept in terms of understanding the performance of estimators, let's return to our $X_i \overset{\text{iid}}{\sim} \text{Cauchy}(\theta)$ example from the previous lecture
- Consider two potential estimators, the sample median $\tilde{\theta}$ and the "one-step" estimator where we solve the likelihood equations using a Taylor series approximation about $\tilde{\theta}$
- Now, it can be shown that

$$\sqrt{n}(\tilde{\theta} - \theta) \xrightarrow{\text{d}} \text{N}(0, \pi^2/4)$$
$$\sqrt{n}(\hat{\theta} - \theta) \xrightarrow{\text{d}} \text{N}(0, 2)$$

Asymptotic relative efficiency

- Since $\pi^2/4 = 2.47 > 2$, we can now appreciate the purpose of the one-step estimator: while both estimates are consistent, the one-step estimator is more efficient
- **Definition:** If $\sqrt{n}(\hat{\theta}_1 - \theta) \xrightarrow{\mathrm{d}} \mathrm{N}(0, \sigma_1^2)$ and $\sqrt{n}(\hat{\theta}_2 - \theta) \xrightarrow{\mathrm{d}} \mathrm{N}(0, \sigma_2^2)$, the *asymptotic relative efficiency* (ARE) of the two estimators is σ_1^2/σ_2^2
- For the Cauchy estimators, the ARE is $2.47/2 = 1.23$
- In other words, the median estimator requires approximately 23% larger sample size than the one-step estimator: we need $n = 123$ observations with the median estimator to obtain the same amount of information that the one-step estimator has with $n = 100$

Asymptotic relative efficiency: Tests

- This idea can be extended to testing as well
- Since the power of any reasonable test tends to 1 as $n \to \infty$, one typically considers $H_0 : \theta = \theta_0$ vs $H_a : \theta = \theta_0 + \Delta/\sqrt{n}$
- In this case, if $\beta_1 \to \Phi(\Delta a_1 - Z_{1-\alpha})$ and $\beta_2 \to \Phi(\Delta a_2 - Z_{1-\alpha})$, where β_i is the power of test i, the asymptotic relative efficiency of the two tests is $(a_1/a_2)^2$
- More abstractly, if two statistical procedures have the same limit as $n_1 \to \infty$ and $n_2 \to \infty$, then the ARE is the limit of the ratio n_1/n_2; the estimation and testing definitions we have given are special cases

Asymptotic relative efficiency: Tests (cont'd)

- For example, when $X_i \overset{\text{iid}}{\sim} \mathrm{N}(\Delta/\sqrt{n}, \sigma^2)$, the one-sample t-test satisfies

$$\beta_1 \to \Phi(\Delta/\sigma - Z_{1-\alpha})$$

while the Wilcoxon signed rank test satisfies

$$\beta_2 \to \Phi\left(\frac{\Delta}{\sigma}\sqrt{\frac{3}{\pi}} - Z_{1-\alpha}\right)$$

- Thus, the ARE is $\pi/3 = 1.05$; when the data follows the normal distribution assumed by the t-test, the Wilcoxon test requires just 5% more data in order to achieve the same power

Additional remarks

- If the distribution is not normal, then there is no upper bound the ARE of these two tests – one can always construct a distribution such that the Wilcoxon approach is that many times more efficient than a t-test
- This example illustrates a common use of efficiency: there is often a desire to develop robust nonparametric or semiparametric methods that make less restrictive assumptions than a parametric likelihood model, and efficiency provides something of a gold standard to compare against

Bayesian efficiency

- We mentioned earlier that maximum likelihood estimation is "optimal" in the sense of being asymptotically efficient, but keep in mind that it is not a unique property – there may be multiple efficient approaches
- For example, Bayesian methods are also asymptotically efficient, as we are now going to see (although in the interest of time, I am going to skip the proof)
- There are several versions of this theorem, as it is a problem that has been tackled by many statisticians throughout the years, beginning with Laplace, but the result is usually called the Bernstein-von Mises theorem

Regularity conditions

- There are “local” versions of the Bernstein-von Mises theorem, but I will present the “global” version concerning the posterior mean in Lehmann’s Theory of Point Estimation, which he attributes to Peter Bickel
- We require the same regularity conditions as in the MLE case:

(B1) Assumptions (A), (B), and (C) from the lecture on likelihood consistency are met

- Since the posterior mean requires integrating over all values of $\boldsymbol{\theta}$, however, it is not enough to require likelihood conditions only on a local neighborhood $\boldsymbol{\Theta}^*$; we need to ensure that the likelihood behaves reasonably even at values of $\boldsymbol{\theta}$ far from $\boldsymbol{\theta}^*$

Regularity conditions (cont'd)

(B2) For all $\epsilon > 0$, there exist $\delta > 0$ such that in the expansion

$$\ell(\boldsymbol{\theta}) = \ell(\boldsymbol{\theta}^*) + (\boldsymbol{\theta} - \boldsymbol{\theta}^*)^\top \mathbf{u}(\boldsymbol{\theta}^*) - \frac{1}{2}(\boldsymbol{\theta} - \boldsymbol{\theta}^*)^\top [\boldsymbol{\mathcal{I}}(\boldsymbol{\theta}^*) + \mathbf{R}(\boldsymbol{\theta}^*)](\boldsymbol{\theta} - \boldsymbol{\theta}^*),$$

the probability of the event

$$\sup_{\|\boldsymbol{\theta} - \boldsymbol{\theta}^*\| < \delta} \left| \frac{1}{n} R_{ij}(\boldsymbol{\theta}) \right| > \epsilon$$

tends to 0 as $n \to \infty$ for all i and j

(B3) For all $\epsilon > 0$, there exist $\delta > 0$ such that the probability of the event

$$\sup_{\|\boldsymbol{\theta} - \boldsymbol{\theta}^*\| \geq \delta} \frac{1}{n} \{\ell(\boldsymbol{\theta}) - \ell(\boldsymbol{\theta}^*)\} < -\epsilon$$

tends to 1 as $n \to \infty$

Bernstein-von Mises Theorem

- Finally, we need two conditions on the prior:

(B4) The prior density $p(\boldsymbol{\theta})$ is continuous and positive for all $\boldsymbol{\theta} \in \boldsymbol{\Theta}$

(B5) The prior expectation exists: $\int \|\boldsymbol{\theta}\| dP(\boldsymbol{\theta}) < \infty$

- **Theorem (Bernstein-von Mises):** Let $\hat{\boldsymbol{\theta}}$ denote the posterior mean. If (B1)-(B5) hold, then

$$\sqrt{n}(\hat{\boldsymbol{\theta}} - \boldsymbol{\theta}^*) \xrightarrow{\text{d}} \text{N}(\mathbf{0}, \boldsymbol{\mathscr{I}}^{-1}(\boldsymbol{\theta}^*)).$$

Remarks

- The novel requirement here is that the prior have positive support for all $\boldsymbol{\theta}$
- This makes intuitive sense: so long as the prior does not rule $\boldsymbol{\theta}^*$ out, then eventually we will have enough data that the likelihood dominates the posterior and agrees with the frequentist likelihood approach
- Obviously, this does not imply that Bayesian and frequentist methods are equivalent (introducing a prior to improve performance at small sample sizes is a major advantage of Bayesian approaches), but it is reassuring to know that given enough data, both schools of thought will agree on an answer if they are working with the same likelihood model